# Short Paper

## Human Pose Estimation Using Consistent Max Covering

Hao Jiang, *Member*, *IEEE*

**Abstract**—A novel consistent max-covering method is proposed for human pose estimation. We focus on problems in which a rough foreground estimation is available. Pose estimation is formulated as a jigsaw puzzle problem in which the body part tiles maximally cover the foreground region, match local image features, and satisfy body plan and color constraints. This method explicitly imposes a global shape constraint on the body part assembly. It anchors multiple body parts simultaneously and introduces hyperedges in the part relation graph, which is essential for detecting complex poses. Using multiple cues in pose estimation, our method is resistant to cluttered foregrounds. We propose an efficient linear method to solve the consistent max-covering problem. A two-stage relaxation finds the solution in polynomial time. Our experiments on a variety of images and videos show that the proposed method is more robust than previous locally constrained methods.

**Index Terms**—Human pose estimation, consistent max covering, linear programming.

## 1 Introduction

Human pose estimation has many important applications in surveillance and human computer interaction. In these applications, rough object silhouettes can often be obtained using simple methods such as background subtraction or color segmentation. It is well known that the segmentation can be used to enhance the body part detection. However, there is little work on how to use the global shape of the foreground to constrain the body part assembly in pose optimization. In this paper, we propose a consistent max-covering scheme to take advantage of this global constraint. The basic idea is that when body parts are assembled, they should cover a region of similar shape to the object foreground and they should follow a valid body plan. We study how we can integrate different cues in a linear formulation which can be efficiently solved to achieve reliable results.

For simple poses with little self-occlusion, a rough body configuration can be extracted by the skeleton operation on clean silhouettes [1]. Machine learning methods have also been studied for pose inference using silhouettes [2], [3]. With sufficient training data, fast pose regression can be achieved. The challenge of machine learning methods is that they have to deal with a large set of training samples and high-dimensional feature space. Dimensionality reduction methods such as probabilistic principal component analysis [4], manifold learning [5], and Gaussian process dynamical model [6] have been used to relieve the dimension explosion problem. These methods are currently used for detecting the restricted classes of poses and to work with clean silhouettes. Shape matching methods have also been widely used in pose estimation [7], [8], [9]. Shape context matching [7] and Chamfer matching [8] are popular methods for finding human poses. Shape matching methods are resistant to clutter, but have high complexity when there are a huge number of exemplars. Hierarchical search trees [8] [18] and locality sensitive hashing [9] have been studied to improve the search efficiency.

Apart from these aforementioned top-down approaches, bottom-up pose estimation methods have also been intensively studied. In the bottom-up methods, body part candidates are first detected and then assembled to fit the image observations and a body plan. This approach is most related to our method. Using silhouettes, an efficient pose estimation method [10] has been studied based on a tree model and posterior sampling. This method is resistant to the foreground clutter. One difficulty of this formulation is the "overcounting" issue, which happens when multiple body parts occupy the same pixel in an image. To relieve the problem, a separate sampling step is used to generate a number of human poses based on the probability estimated in the tree inference stage. Tree structure models can be learned adaptively [26] to achieve reliable results. Other stochastic searching methods [11], [12] have also been used in pose optimization.

Nontree models can be used to enforce tighter structural constraints. A widely used scheme is to include the pairwise constraints between pair of arms and legs to penalize overlapping body parts. These constructions introduce cycles into the body part relation graph; the inference on such loopy graphs is generally NP-hard. Different methods are proposed to solve the hard optimization problem. A branch and bound method [19] is proposed to obtain an exact solution. More efficient approximation methods such as convex programming [13], [15] or belief propagation (BP) [16] have also been proposed to tackle the problem. These nontree methods use local pairwise constraints and have no mechanisms to control the global structure. For complex poses, local constraint is not sufficient to resolve the ambiguity in pose estimation. Without a global cue, it is hard to decide whether two arms or legs should be assigned to the same location or be apart. The uniform penalty used in traditional nontree methods relieves the problem, but it also introduces undesired penalty to truly overlapping body parts.

In this paper, we follow the body part assembly scheme and propose a novel consistent max-covering method for human pose estimation [23]. We assume that a rough foreground potential map is available. Differently from previous locally constrained methods, our method incorporates global object shapes into pose optimization. In our formulation, pose estimation becomes the problem of covering an object foreground map with a set of body part tiles: They maximally cover the object foreground, match image local appearance, and are consistent in terms of the body linkage plan and other symmetry constraints. Max covering with the consistency constraint is denoted as *consistent max covering*. It introduces high-order relations among all of the body parts since each body part may influence others when forming a covering. The high-order body part correlation is essential for complex pose estimation when self-occlusions or other part interactions occur. This method also solves the "overcounting" problem. By encouraging the body parts to cover the object foreground, we remove the undesired penalty on the truly overlapping body parts in pose estimation. In this paper, we formulate the consistent max covering as a linear optimization and we propose an efficient solution using a two-stage relaxation. By incorporating multiple cues in pose estimation, the proposed method is resistant to occlusion and works with low-quality foreground estimation or soft object foreground mask. Our experiments on a variety of images and videos show that the proposed method is robust and efficient in human pose estimation.

• *The author is with the Computer Science Department, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467. E-mail: hjiang@cs.bc.edu.*

*Manuscript received 27 May 2010; revised 30 Nov. 2010; accepted 10 Apr. 2011; published online 28 Apr. 2011.*
*Recommended for acceptance by P. Felzenszwalb.*
*For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-05-0412.*
*Digital Object Identifier no. 10.1109/TPAMI.2011.92.*

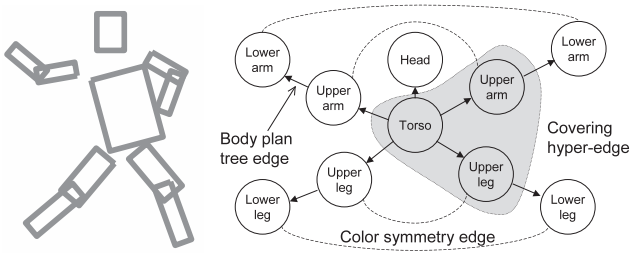


Fig. 1. Left: 10-part body model. Right: Relation graph of body parts. The gray area shows the example of a hyperedge.

## 2 Related Work

Object foreground estimation has been used in different ways to improve human pose estimation. Ferrari et al. [20] use rough foreground segmentation to improve the local body part detection for human upper body pose estimation. Mori and Malik [7] use superpixels to find salient body parts in images. Ren et al. [13] propose another approach for human pose estimation in static images using superpixels. In this method, an integer program is formulated and relaxed to linear program for pose optimization. Ramanan [14] proposes an effective method to train local body part detectors using a soft segmentation; a tree structure model is used in the optimization. Johnson and Everingham [21] propose another method to use object segmentation to enhance the local body part detection; the pose estimation follows the tree structure inference method in [14]. In these methods, the segmentation information has not yet been fully used; the object foreground estimation is used as a means to improve local part detection, but not incorporated in the human pose optimization. In this paper, we show how we can seamlessly merge the object foreground constraint in pose optimization.

Linear methods have been receiving a lot of interest in recent years for solving computer vision problems. These methods can be relaxed and globally optimized. Previous linear methods for pose estimation use pairwise or lower order constraints. Ren et al. [13] use linear programming (LP) relaxation for pose estimation with pairwise constraints. A linear method that uses pairwise and triorder exclusion constraints in an augmented tree model is studied in [15]. Differently from these approaches, our proposed model is able to enforce higher order constraints, which are hard to implement using previous formulations.

## 3 Consistent Max Covering for Pose Estimation

Given an object foreground map from color segmentation or background subtraction, pose estimation can be simulated as a jigsaw puzzle problem. In the following, we show how pose estimation can be formulated as consistent max covering and how we can solve it using an efficient linear method.

### 3.1 Body Model and Part Detection

We use the widely used 10-part body model which contains head, torso, upper arms, lower arms, upper legs, and lower legs. Each body part is represented as a rectangle. Our method can also be easily extended to more complex body part shapes. The 10-part body model and the part relation graph are shown in Fig. 1. In our model, the basic body plan follows a tree structure. Apart from the interactions between neighboring body parts, consistent max-covering formulation introduces hyperedges linking the body parts that cover the same foreground region, and the edges that constrain symmetry body parts. The basic body plan tree is rooted at the torso and has directional edges. The other two kinds of edges are nondirectional.

Similarly to other bottom-up pose estimation methods, we first locate potential body part candidates in target images so that we can use them in the consistent max covering. We assume a fixed object scale in this paper; when the foreground can be accurately extracted, the object scale can be estimated and normalized automatically. We use simple box detectors to find the body part candidates on the edge map of the target image. Chamfer matching is used to match body part templates to the target edge map at different locations and rotations. Nonminimum suppression is then used to locate the body part candidates. Since we have a rough foreground map, the body part candidates can be further pruned: We only keep the candidates whose average foreground potentials are greater than a threshold. The linear combination of the local Chamfer matching cost and the foreground covering cost is associated with each body part candidate as the local matching cost. A body part candidate is represented as a rectangle with a start side and an end side.

### 3.2 Consistent Max Covering: The Overview

Each body part candidate covers some pixels in the object foreground. Intuitively, the body part tiles should cover foreground pixels as much as possible in a consistent manner.

Assume that we obtain a foreground estimation, and a floating-point number from 0 to 1 is related to each image pixel to indicate the foreground potential. The higher the potential, the more likely it is that the pixel belongs to the object foreground. We denote the foreground map as $f_{x,y}$. The consistent max covering can be formulated as the following optimization problem:

$$\begin{aligned} &\max_{\mathcal{C}} \left\{ \sum_{(x,y)\in I} r_{x,y} - \alpha M(\mathcal{C}) - \beta P(\mathcal{C}) - \gamma S(\mathcal{C}) \right\} \\ &s.t. \quad r_{x,y} = f_{x,y}, \text{ if } (x,y) \text{ is covered by parts, else } 0 \\ &\quad \mathcal{C} \text{ is a body part covering,} \end{aligned} \tag{1}$$

where $r_{x,y}$ is the covered potential at pixel $(x, y)$ with the current body part covering $\mathcal{C}$: If the pixel $(x, y)$ in the foreground $I$ is covered by the body parts, $r_{x,y}$ takes value $f_{x,y}$ and otherwise 0. Therefore, the first term in the objective function equals the overall potential covered by all the body parts. The second term $M(\mathcal{C})$ is the cost of matching the body parts to local image features. The third term $P(\mathcal{C})$ is the degree of the body part configuration following a human body plan. The last term $S(\mathcal{C})$ penalizes the color difference of symmetrical body parts: If the symmetrical parts, e.g., upper arms, have large color difference, $S$ has a large value. We reverse the sign of the last three terms so that they are minimized. $\alpha$, $\beta$, and $\gamma$ are positive constants to control the weight among the energy terms. This optimization thus tends to find a consistent max covering on the object foreground.

Consistent max covering in (1) is a combinatorial search problem. We need to find a body part configuration to optimize the objective while satisfying the constraints. It is generally NP-hard because of the loopy part relations introduced by the covering terms and symmetrical relation terms. The large number of feasible body part configurations makes naive exhaustive search infeasible and, for such a problem, the greedy method is not sufficient. We use a global search method to tackle this problem. In the following, we propose an efficient linear solution.

### 3.3 Linear Formulation

We linearize the consistent max-covering optimization in (1) and obtain a mixed integer linear program. It can be further relaxed into a much simpler linear program for efficient solution.

#### 3.3.1 Foreground Covering Potential

In (1), the covering term in the objective function is $\sum_{(x,y)\in I} r_{x,y}$, which equals the total potential covered by all the body parts.

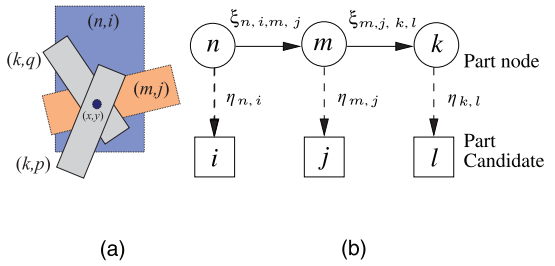


Fig. 2. (a) Body part covering example in which part candidates $(n,i)$, $(m,j)$, $(k,p)$, and $(k,q)$ cover pixel $(x,y)$ in the foreground map. (b) Definitions of edge variables and node variables.

Recall that $r$ represents the covering potential at a pixel. We now explicitly express how the variable $r$ is related to the choice of body part locations.

We introduce a binary indicator variable $\eta_{n,i}$, which is 1 if body part $n$ selects target candidate $i$, otherwise 0. $\eta_{n,i}$ is therefore a node variable that corresponds to each node of the body graph. Since each body part only has one target location, we need to make sure

$$\sum_{i \in T(n)} \eta_{n,i} = 1, \quad \forall n \in V,$$

where $T(n)$ is the target candidate set of body part $n$; $V$ is the set of all body parts; we also use $V$ to denote the node set of the body graph. We are now ready to specify the covering term using the node variable $\eta$. We introduce the following constraint for the covering variable $r$:

$$\sum_{\forall (n,i) \text{ covers } (x,y)} \eta_{n,i} \geq r_{x,y},$$

in which $(n,i)$ denotes the $i$th candidate for body part $n$. We further need to bound $r_{x,y}$ to be a nonnegative number that can be as large as the foreground potential

$$0 \leq r_{x,y} \leq f_{x,y}.$$

For $r_{x,y}$ not covered by any candidate tiles, it is set to zero. In an example shown in Fig. 2a, there are total four body part candidates $(n,i)$, $(m,j)$, $(k,p)$, and $(k,q)$ covering the pixel $(x,y)$. The constraint for $r_{x,y}$ is therefore $r_{x,y} \leq \eta_{n,i} + \eta_{m,j} + \eta_{k,p} + \eta_{k,q}$ and $0 \leq r_{x,y} \leq f_{x,y}$.

It is not hard to verify that with the above formulation, $r_{x,y}$ equals the covering potential at pixel $(x,y)$. If at least one body part covers the foreground pixel $(x,y)$, to maximize the objective function in (1), $r_{x,y}$ should equal $f_{x,y}$, recalling that body part indicator variable $\eta$ is 0 or 1 and $f_{x,y}$ is between 0 and 1; if $(x,y)$ is not covered by any part tiles, $r_{x,y}$ will be 0 because of the upper bound constraint. Therefore, given such constraints, $\sum_{(x,y) \in I} r_{x,y}$ is indeed the overall covering potential of the body parts.

### 3.3.2 Image Matching Cost

Apart from the object foreground covering cost, assigning each body part to a target location involves an image matching cost. In this paper, the cost is the linear combination of the Chamfer matching cost and the local covering cost. Local covering cost equals 1 minus the average covering potential. With the binary node assignment variable $\eta$, the total image matching cost can be linearized as

$$M = \sum_{n \in V, i \in T(n)} c_{n,i} \cdot \eta_{n,i}.$$

Here, $c_{n,i}$ is the image matching cost of part $n$ at location $i$.

### 3.3.3 Body Plan Energy

A good body part assignment should have neighboring body parts linked together: The end points of consecutive body parts are close, and the connected limbs have similar orientations. To linearize the spatial consistency function $P(.)$ in (1), we introduce an edge indicator variable $\xi_{n,i,m,j}$ for each directional tree edge $\langle n,m \rangle$; $\xi_{n,i,m,j}$ is 1 if body part $n$ selects candidate $i$ and body part $m$ takes candidate $j$. The definitions of the edge variable $\xi$ and node variable $\eta$ are illustrated in Fig. 2b.

Using edge assignment variable $\xi$, the total spatial consistency cost is linearized as

$$P = \sum_{\langle n,m \rangle \in E, i \in T(n), j \in T(m)} h_{n,i,m,j} \cdot \xi_{n,i,m,j},$$

where $E$ is the tree edge set and the coefficient $h_{n,i,m,j}$ is defined as

$$h_{n,i,m,j} = \begin{cases} a e_{n,i,m,j}, & \text{if } n \text{ is torso}, \\ a d_{n,i,m,j} + b \sin^2\left(\frac{\theta_{n,i} - \theta_{m,j}}{2}\right), & \text{otherwise}. \end{cases}$$

Here, $e_{n,i,m,j}$ is the euclidean distance between a proper end of a torso candidate and the start point of an upper body part candidate; $d_{n,i,m,j}$ is the euclidean distance between the end point of part $n$ at location $i$ and the start point of part $m$ at location $j$; $\theta_{n,i}$ is the angle of part $n$ at location $i$ and $\sin^2\left(\frac{\theta_{n,i} - \theta_{m,j}}{2}\right)$ penalizes large angle difference for consecutive body parts; $a$ and $b$ are positive weight constants.

The node variable $\eta$ and the edge variable $\xi$ are dependent. The pairwise edge assignment has to be consistent with the node assignment: Each body part appearing in different pairs must have a unique assignment of the target candidate. To enforce the assignment consistency, $\forall \langle n,m \rangle \in E$, we let

$$\eta_{n,i} = \sum_{j \in T(m)} \xi_{n,i,m,j}, \quad \eta_{m,j} = \sum_{i \in T(n)} \xi_{n,i,m,j}.$$

Recall that $T(m)$ is the set of body part candidates for part $m$. The above constraints imply that $\sum_{j \in T(m)} \xi_{n,i,m,j} = \sum_{l \in T(k)} \xi_{n,i,k,l}$, $k \neq m$, and $\sum_{i \in T(n)} \xi_{n,i,m,j} = \sum_{l \in T(k)} \xi_{m,j,k,l}$. This enforces the assignment consistency at common nodes in the tree.

### 3.3.4 Color Consistency

To linearize the color difference term $S$ in (1), we use the $L_1$ norm to compute the color difference of two body parts so that we can use a standard linear programming auxiliary variable trick [17]. Let $H$ be the set of symmetrical body part pairs. Term $S$ can be linearized as

$$S = \sum_{\{n,m\} \in H} \sum_{k=1}^{3} \left( g^+_{n,m,k} + g^-_{n,m,k} \right),$$

where $g^+_{m,n,k}$ and $g^-_{m,n,k}$ are nonnegative auxiliary variables. We use three color channels $k = 1..3$. The nonnegative auxiliary variables are constrained by the color difference at each channel:

$$g_{n,k} - g_{m,k} = g^+_{n,m,k} - g^-_{n,m,k}, \quad k = 1..3, \quad \forall \{n,m\} \in H.$$

Here, $g_{n,k}$ is the color of body part $n$ at channel $k$. The color of a body part can be computed using the node assignment indicator variable $\eta_{n,i}$:

$$g_{n,k} = \sum_{i \in T(n)} l_{n,i,k} \cdot \eta_{n,i},$$

and $l_{n,i,k}$ is the average color of the candidate covering region $i$ for body part $n$ at channel $k$. It is easy to verify that at least one variable in the pair of $g^+_{n,m,k}$ and $g^-_{n,m,k}$ will become 0 when the objective function is optimized; otherwise, we can zero one of them and obtain a better solution by subtracting the two variables with

the smaller one of them. Therefore, when the objective function is optimized, $g^+_{n,m,k} + g^-_{n,m,k} = |g_{n,k} - g_{m,k}|$, and $S$ is the $L_1$ color distance between symmetrical body parts. Note that the color symmetry term is a regularization term in the objective function. The constraint of color is therefore soft, which permits occasionally large discrepancy of colors on symmetrical body parts.

#### 3.3.5 Relaxation and Two-Stage Approximation

The consistent max-covering optimization is a mixed integer program with binary variables $\xi$ and $\eta$, and continuous variables $r$ and $g$. Directly solving the mixed integer program has high complexity. We relax it into the following linear program:

$$
\begin{aligned}
\max \Bigg\{ & \sum_{(x,y)\in I} r_{x,y} - \alpha \sum_{n\in V, i\in T(n)} c_{n,i}\cdot \eta_{n,i} \\
& - \beta \sum_{\langle n,m\rangle \in E, i\in T(n), j\in T(m)} h_{n,i,m,j}\cdot \xi_{n,i,m,j} \\
& - \gamma \sum_{\{n,m\}\in H} \sum_{k=1}^{3} \left(g^+_{n,m,k} + g^-_{n,m,k}\right) \Bigg\} \\
s.t. \quad & \eta_{n,i} = \sum_{j\in T(m)} \xi_{n,i,m,j}, \quad \eta_{m,j} = \sum_{i\in T(n)} \xi_{n,i,m,j}, \\
& \xi \geq 0, \quad \forall \langle n,m\rangle \in E \\
& \sum_{i\in T(n)} \eta_{n,i} = 1, \eta_{n,i} \geq 0, i \in T(n), \quad \forall n \in V \\
& g_{n,k} - g_{m,k} = g^+_{n,m,k} - g^-_{n,m,k}, \\
& g^+_{n,m,k},\ g^-_{n,m,k} \geq 0, \quad \forall \{n,m\} \in H \\
& g_{n,k} = \sum_{i\in T(n)} l_{n,i,k}\cdot \eta_{n,i}, \quad k = 1..3, \quad \forall n \in V \\
& \sum_{\forall (n,i) \text{ covers } (x,y)} \eta_{n,i} \geq r_{x,y}, \\
& 0 \leq r_{x,y} \leq f_{x,y}, \quad \forall (x,y) \in I,
\end{aligned}
$$

where $\xi$ and $\eta$ are relaxed into continuous variables in $[0, 1]$. If we do not include the covering constraint term and color symmetry term, the linear program on the tree structure body plan is equivalent to the integer program and it can be solved efficiently using dynamic programming (DP). The nontree structure of consistent max covering complicates the solution. Its relaxation does not directly yield integer solutions for node variables $\eta$. Rounding by selecting the largest $\eta$ for each body part yields poor results. Fortunately, using the interior method, its solution almost always contains very few large $\eta$. We threshold $\eta$ to zero out most variables. The typical threshold is 0.001. Similarly to the approximation trick in [15], we can further construct a small mixed integer program by only including the target candidates corresponding to the nonzero $\eta$. The small mixed integer program can be directly solved using an exhaustive enumeration or a more efficient branch and bound method. Since the first step eliminates a large number of body part candidates, the second exhaustive search step can be quickly solved.

The average complexity of a linear program is roughly linear to the number of constraints and logarithm to the number of variables [17]. This simplex method heuristic applies to our model for which the primal-dual interior point method is almost always faster than the simplex method for different problem sizes. The number of edge variables of the proposed linear program is proportional to the square of the number of target candidates $n$; the number of foreground variables equals the number of foreground pixels $m$. The number of constraints is in the same order as the number of variables. The linear program thus has $O((n^2 + m)\log(n^2 + m))$ average complexity. We can further speed up the linear program by heuristics. The neighboring body parts only accept quite limited set of candidates: The pair of candidates too far away can be pruned. Using such a trick, the number of edge variables and related constraints can be greatly reduced. The number of foreground variables and constraints can also be reduced by using a coarser representation of the foreground map: Instead of corresponding to each pixel, the foreground variable and constraint correspond to each region. Due to slow variation of the foreground map, downsampling will not degrade the performance. Typically, we extract 100 torso candidates, 200-500 limb and head candidates, and use 2,500-5,000 foreground variables to represent foreground map regions. In a 2.8 GHz Linux machine, the linear program takes less than 20 seconds to converge. We need a few parameters in our formulation to control the weight of different energy terms. In this paper, we simply set the parameters by trial and error and fix them in all of the experiments. A systematic learning approach can be used to fine-tune these parameters on ground truth data set using exhaustive search or generic nondifferentiable local optimization methods.

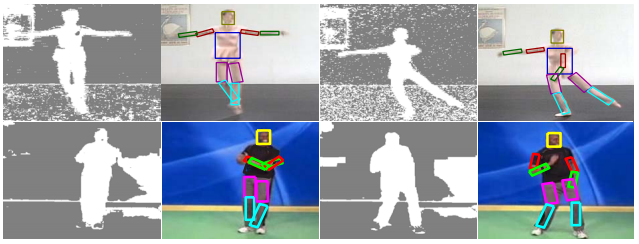

Fig. 3. Pose estimation with cluttered foreground. The gray-scale images on the first and third columns are foreground maps; gray indicates potential 0 and white 1. The pose estimation results are overlapped on the corresponding color images.

### 3.4 Exclusion or Covering: A Comparison

In [15], we use exclusion constraints to penalize overlapping body parts. This constraint prevents different body parts from occupying the same spatial location. To implement the exclusion constraint, we first find all of the body part candidates that conflict with $(n, i)$ and we denote the set as $A_{n,i}$ and then we enforce

$$\eta_{n,i} + \sum_{(m,j)\in A_{n,i}} \eta_{m,j} \leq 1, \quad i \in T(n).$$

With such a constraint, if $(n, i)$ is assigned to part $n$, other body parts cannot select any candidates in the set that conflicts with $(n, i)$.

Consistent max covering and exclusion are two different ways to "spread out" the body part assignment. They seem to achieve similar effect, but in fact they behave quite differently: Consistent max covering does not penalize truly overlapping body parts, instead it encourages body parts to cover the object foreground; body part exclusion penalizes all the overlapping patterns even though the parts should really overlap. We compare the performance of these two linear schemes in the experimental section.

## 4 Experimental Results

In this section, we evaluate the consistent max-covering method on human pose estimation and compare it with different approaches.

Fig. 3 illustrates the proposed method's resistance to cluttered foreground. Fig. 4 shows another test in which the foreground map is corrupted by a higher level of structure noise. The results degrade mildly even though there is a substantial amount of clutter in both the object foreground and background. In the proposed method, max covering is a soft constraint; it therefore allows rough foreground estimations. When the foreground map is poor and every pixel's foreground potential approaches 0.5, the optimization is equivalent to pose detection with uniform penalty on overlapping body parts.

Fig. 4. The proposed method still works well when the foreground map is corrupted by structured noise. The foreground corresponds to the "brighter" pixels in the images. The result degrades mildly; in the 710 body part detections, we have 16 more part detection errors on the noisy foreground maps than those on the clean foregrounds.

Fig. 5. Compare with pose estimation using max covering which ignores the consistency constraints. From left to right: Input image, foreground map, pose estimation using max covering, and the result of consistent max covering.

Sample results of pose estimation using the proposed method are shown in Fig. 6. The test data include six test video sequences: two ballet sequences, two lab sequences, and two videos from YouTube. The ballet sequences include complex movements and body part self-occlusions. In the YouTube sequences and the lab sequences, actors wear baggy clothes and perform complex movements. YouTube videos also have low image quality due to heavy compression. There are 1,464 images in total. The proposed method robustly detects body poses in the test sequences. In the following, we compare the proposed method with different competing approaches.

### 4.1 Comparison with Different Variations

We first compare the proposed method with some of its variations. Fig. 5 illustrates the results of max covering and consistent max covering. Max covering maximizes the covering potential while ignoring other consistency constraints. As expected, max covering may generate a body part covering that does not resemble a human body plan. Consistent max covering is necessary to obtain good results.

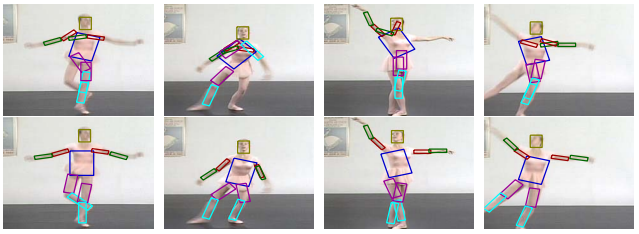

Fig. 7. Compare with DP. The first row is the result of DP and the second row shows how the proposed method improves the result.

TABLE 1
Average Number of Errors Per Frame in Pose Estimation
(B1-2, L1-2, T, F Indicate Ballet, Lab, Taichi, and Fitness Videos)

| | B1 | B2 | L1 | L2 | T | F |
|---|---|---|---|---|---|---|
| This Paper | 0.46 | 0.74 | 0.90 | 0.77 | 1.29 | 1.40 |
| DP | 3.91 | 3.47 | 6.01 | 3.20 | 3.30 | 3.68 |
| Simple LP | 0.97 | 1.43 | 1.46 | 1.27 | 2.31 | 2.46 |
| Tree-I [14] | 2.80 | 2.48 | 1.29 | 1.09 | 2.40 | 1.56 |
| Tree-II [22] | 1.38 | 1.24 | 0.95 | 1.21 | 2.18 | 1.42 |
| Non-Tree [15] | 1.30 | 1.26 | 1.51 | 1.46 | 2.19 | 2.08 |
| BP | 1.95 | 1.80 | 3.09 | 3.37 | 2.96 | 3.33 |

We proceed to test another variation of the linear method in which we keep only the local part matching cost and the tree structure spatial consistency constraint. Ignoring the covering energy, this is in fact the formulation which can be solved by DP [10]. The proposed relaxation method is exact and equivalent to DP in this case. Without global constraints, we expect more errors. Fig. 7 shows how the proposed method improves the result over DP. The comparison is based on both visual inspection and comparing with the ground truth labeling. The qualitative comparison of average detection errors is listed in Table 1. With ground truth data, we further define the pose score of each detection as the ratio of the overlapping area of limb detection with the ground truth region to the overall limb foreground. The most easily detected torso is not counted so the measurement is more sensitive. A perfect detection has the pose score of 1. In real detections, pose score is from 0 to 1 and a higher pose score indicates superior performance. Fig. 9 shows the normalized histograms of per-frame pose scores. The ideal pose detector should have a single peak of 1 at pose score 1 and vanishes anywhere else. A good real detector has a pose score histogram focused on the right side and has a steep rise on the left

Fig. 6. Pose estimation using consistent max covering for the six test videos. `Ballet` has complex movements, self-occlusion, fast motion, and interleaving effects; `lab` involves baggy clothes, cluttered background and foreground; actors wear baggy clothes in the `taichi` and `fitness` sequences from YouTube. We use simple color segmentation to extract foreground maps for all the sequences except `lab-II` for which we use background subtraction. Row 1: Sample results for 301-frame `ballet-I` and 314-frame `ballet-II`. Row 2: Sample results for 186-frame `lab-I`. Row 3: Sample results for 276-frame `lab-II`. Row 4: Sample results for 303-frame `taichi`. Row 5: Sample results for 84-frame `fitness`.

TABLE 2
Average Pose Scores

| | B1 | B2 | L1 | L2 | T | F |
|---|---|---|---|---|---|---|
| This Paper | 0.85 | 0.83 | 0.81 | 0.82 | 0.72 | 0.81 |
| DP | 0.63 | 0.64 | 0.44 | 0.55 | 0.61 | 0.65 |
| Simple LP | 0.79 | 0.77 | 0.78 | 0.78 | 0.65 | 0.70 |
| Tree-I [14] | 0.61 | 0.62 | 0.76 | 0.80 | 0.57 | 0.77 |
| Tree-II [22] | 0.77 | 0.74 | 0.78 | 0.76 | 0.70 | 0.78 |
| Non-Tree [15] | 0.79 | 0.76 | 0.70 | 0.67 | 0.57 | 0.71 |
| BP | 0.78 | 0.75 | 0.59 | 0.65 | 0.61 | 0.67 |

TABLE 3
Percentage of Detections with Pose Score $< 0.7$

| | B1 | B2 | L1 | L2 | T | F |
|---|---|---|---|---|---|---|
| This Paper | 2.7 | 6.7 | 4.3 | 3.2 | 41.6 | 4.8 |
| DP | 61.1 | 57.0 | 96.8 | 85.5 | 72.9 | 55.9 |
| Simple LP | 18.9 | 23.9 | 10.8 | 17.8 | 64.4 | 52.4 |
| Tree-I [14] | 83.1 | 70.1 | 22.6 | 13.8 | 76.9 | 14.3 |
| Tree-II [22] | 22.6 | 24.8 | 16.7 | 22.8 | 43.9 | 7.1 |
| Non-Tree [15] | 19.2 | 25.8 | 43.5 | 56.2 | 76.9 | 36.9 |
| BP | 17.9 | 27.1 | 86.0 | 65.9 | 74.6 | 52.4 |

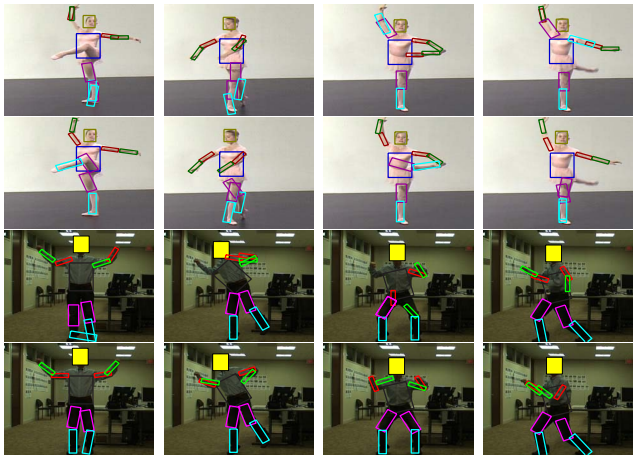

Fig. 8. Compare with a simpler linear formulation (Simple LP) which uses only node variables. The odd rows are the results of Simple LP and the even rows show the results of the proposed method.

side. As shown in Fig. 9, the proposed method's result is substantially better than that of DP. The average pose score in Table 2 confirms the observation. Large detection errors are characterized by pose scores less than 0.7. As shown in Table 3, the number of large errors of the proposed method are often 10 times smaller than those of DP.

We further compare the proposed method with a closely related linear programming formulation. If the body part end point distance is measured with $L_1$ norm, we can use the auxiliary variable trick to construct a simpler linear formulation that includes only node variables. In this restrictive case, the proposed method and the simple LP formulation are equivalent under integer constraints. But the relaxation has a large difference. As shown in Fig. 8, the proposed method yields better results for the challenging cases. The comparison of the two methods using all six test videos is summarized in Fig. 9 and Tables 1, 2, and 3. The proposed method is consistently better than the simple LP over all the test cases with about half of the per-frame errors.

### 4.2 Comparison with Other Pose Detection Methods

Our previous tests show that the proposed method is indeed better than its related variations. The question is, does it generate better results than locally constrained methods? We compare the proposed method with the inference method using tree structure

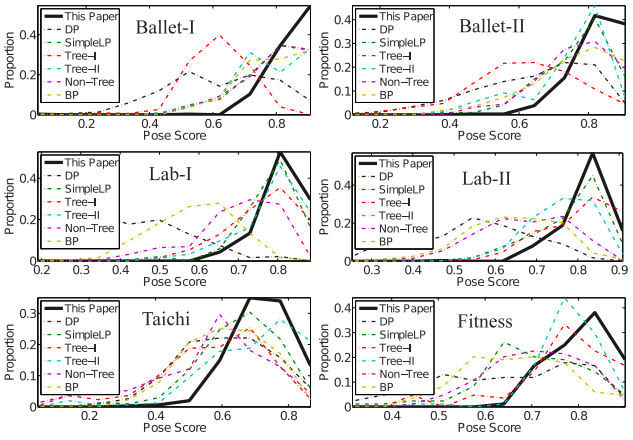


Fig. 9. Per-frame pose score distributions. Good performance is characterized by a large portion of a curve in the high score range.

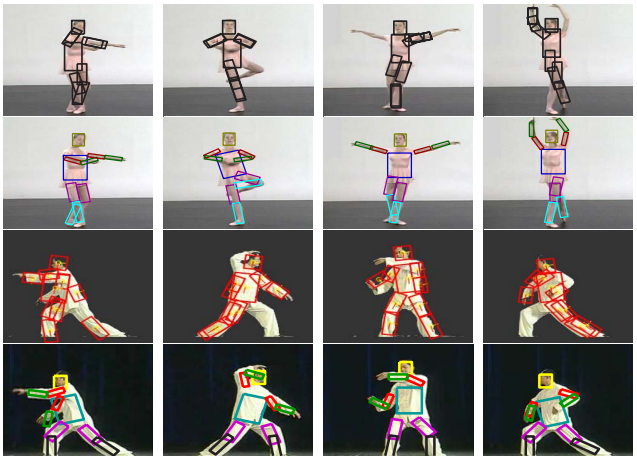

Fig. 10. Compare with the tree inference method one (Tree-I) [14] and two (Tree-II) [22]. Rows 1 and 3 show the results of tree-I and tree-II. Rows 2 and 4 show the results of the proposed method.

[14] and a recent tree method using stronger part detectors [22]. We run the code with these papers on our data. For fair comparison, we use foreground mask to partially eliminate the background clutter before using the code for body part detection. The two tree methods use stronger body part detectors than the simple bar detectors in this paper. As shown in Fig. 10, tree methods sometimes miss detecting body parts because they do not incorporate the global body shape; the proposed method gives better results. The pose score histograms are shown in Fig. 9, and Tables 1, 2, and 3 show more statistics of these methods. The proposed method gives better results over all the test cases.

We further compare the proposed method with a nontree method [15] and a BP-based method. The BP method is implemented using libDAI [25]. These methods use local constraints to penalize the overlapping body part detections to form nontree model graphs. The results are shown in Figs. 9 and 11, and Tables 1, 2, and 3. The proposed method still yields much better results than the two competing nontree methods. As shown in Fig. 6, a few more errors occur in the taichi and fitness sequences. This is mostly due to the simple body part detector used in this paper. Weak image edges may result in error local matching costs associated with body part candidates. A stronger body part detector will further improve the results.

### 4.3 Test on More Ground Truth Data

We test the proposed method on more ground truth data including the walking sequence in [16], the CMU motion capturing data, and the videos from the HumanEva data set [24].

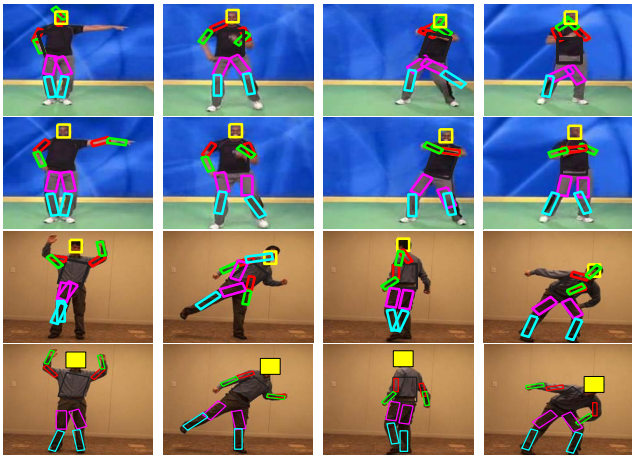

Fig. 11. Compare with the nontree method [15] (row 1) and BP (row 3). The odd rows are the results using the competing methods and the even rows show results of the proposed method.

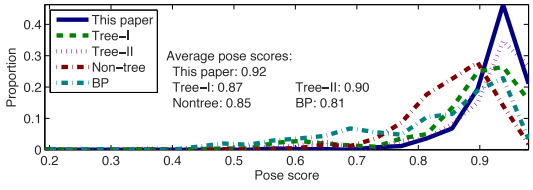


Fig. 12. Comparison on the synthetic data. One thousand poses are generated using the CMU motion capturing data. We compare the pose score of the proposed method with the tree inference method [14], [22], the nontree method [15], and the BP method. Higher pose score indicates better performance.

We compare with the method in [16] on the 50-frame walking sequence. Our mean detection error is 9.8 pixels, compared to 10.3 reported in [16]. Our experiment setting is looser than that in [16]: We do not assume the occlusion order and we do not use any training data. Our body model is an average person model. We simply manually set the scale. Our method is thus able to give more reliable results than [16].

We further test the proposed method on synthetic data generated from the CMU motion capturing data set. One thousand poses are randomly selected from the data set and the front views are rendered. We apply the proposed method, the two-tree inference methods [14], [22], the nontree method [15], and the BP method to the synthetic images to detect human poses. The detections are compared with the ground truth. Fig. 12 shows the comparison result of the proposed method with the tree methods, the nontree method, and the BP method. The normalized histograms of the pose scores of the five methods indicate that the proposed method has the best performance. The average pose score comparison confirms the observation.

We also use the HumanEva [24] data set in the comparison. The three most cluttered video sequences from camera view one are selected in the comparison. Background subtraction is used to obtained the object foreground. Due to the very cluttered background, the object foreground estimation is quite noisy. We roughly normalize the object scale and then optimize the pose estimation. Fig. 13 compares the pose scores for the three test videos. The proposed method has the best performance. The visual inspection result conforms to the quantitative analysis.

## 5 Conclusion

We propose a novel consistent max-covering scheme for human pose estimation. The proposed method seamlessly combines different cues such as edges, color symmetry, body linkage plan,

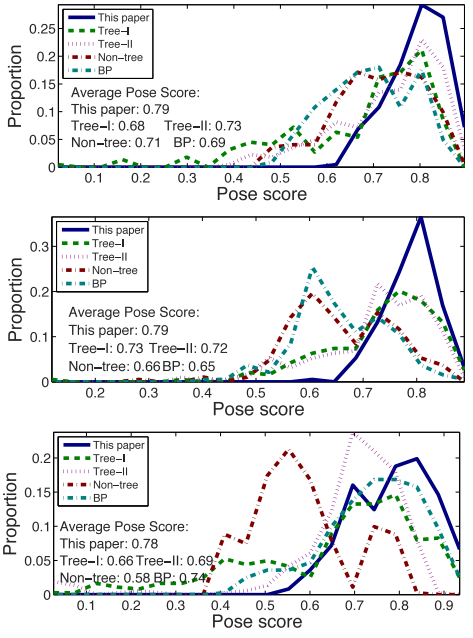


Fig. 13. Comparison on the HumanEva data set [24]. The walking, boxing, and jogging sequences from camera one are used in the experiment. Row 1: walking, Row 2: boxing, Row 3: jogging. There are a total of 1,182 frames.

and finds a consistent max covering of the object foreground map using body part polygons. It introduces high-order correlations among multiple body parts and greatly improves the performance of pose estimation for complex movements. We devise a linear formulation and an efficient relaxation method to solve consistent max covering. Experiments on challenging images and videos show that the proposed method is robust and efficient. We believe that this method is useful for many applications including automatic surveillance and human movement analysis.

## Acknowledgments

This work is supported by US National Science Foundation (NSF) Grant 1018641.

## References

[1] X. Bai and W.Y. Liu, "Skeleton Pruning by Contour Partitioning with Discrete Curve Evolution," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 449-462, Mar. 2007.

[2] R. Rosales and S. Sclaroff, "Inferring Body Pose without Tracking Body Parts," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[3] A. Agarwal and B. Triggs, "3D Human Pose from Silhouettes by Relevance Vector Regression," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.

[4] K. Grauman, G. Shakhnarovich, and T. Darrell, "Inferring 3D Structure with a Statistical Image-Based Shape Model," *Proc. IEEE Ninth Int'l Conf. Computer Vision*, 2003.

[5] A. Elgammal and C.S. Lee, "Inferring 3D Body Pose from Silhouettes Using Activity Manifold Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.

[6] J.M. Wang, D.J. Fleet, and A. Hertzmann, "Gaussian Process Dynamical Models for Human Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 283-298, Feb. 2008.

[7] G. Mori and J. Malik, "Estimating Human Body Configurations Using Shape Context Matching," *Proc. Seventh European Conf. Computer Vision*, 2002.

[8] D.M. Gavrila, "A Bayesian, Exemplar-Based Approach to Hierarchical Shape Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1408-1421, Aug. 2007.

[9] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast Pose Estimation with Parameter Sensitive Hashing," *Proc. IEEE Ninth Int'l Conf. Computer Vision*, 2003.

[10] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial Structures for Object Recognition," *Int'l J. Computer Vision*, vol. 61, no. 1, 2005.

[11] S. Ioffe and D.A. Forsyth, "Probabilistic Methods for Finding People," *Int'l J. Computer Vision*, vol. 43, no. 1, June 2001.
[12] M.W. Lee and I. Cohen, "Proposal Maps Driven MCMC for Estimating Human Body Pose in Static Images," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.
[13] X. Ren, A. Berg, and J. Malik, "Recovering Human Body Configurations Using Pairwise Constraints between Parts," *Proc. IEEE 10th Int'l Conf. Computer Vision*, 2005.
[14] D. Ramanan, "Learning to Parse Images of Articulated Objects," *Neural Information Processing Systems*, 2006.
[15] H. Jiang and D.R. Martin, "Global Pose Estimation Using Non-Tree Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.
[16] L. Sigal and M.J. Black, "Measure Locally, Reasoning Globally: Occlusion-Sensitive Articulated Pose Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.
[17] V. Chvátal, *Linear Programming*. W.H. Freeman and Company, 1983.
[18] H. Ning, W. Xu, Y. Gong, and T.S. Huang, "Discriminative Learning of Visual Words for 3D Human Pose Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.
[19] T. Tian and S. Sclaroff, "Fast Globally Optimal 2D Human Detection with Loopy Graph Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.
[20] V. Ferrari, M.M. Jimenez, and A. Zisserman, "Pose Search: Retrieving People Using Their Pose," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.
[21] S. Johnson and M. Everingham, "Combining Discriminative Appearance and Segmentation Cues for Articulated Human Pose Estimation," *Proc. IEEE Int'l Workshop Machine Learning for Vision-Based Motion Analysis*, 2009.
[22] M. Andriluka, S. Roth, and B. Schiele, "Pictorial Structures Revisited: People Detection and Articulated Pose Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.
[23] H. Jiang, "Human Pose Estimation Using Consistent Max-Covering," *Proc. IEEE 12th Int'l Conf. Computer Vision*, 2009.
[24] HumanEva Data Set, http://vision.cs.brown.edu/humaneva, 2011.
[25] libDAI, http://people.kyb.tuebingen.mpg.de/jorism/libDAI, 2011.
[26] B. Sapp, C. Jordan, and B. Taskar, "Adaptive Pose Priors for Pictorial Structures," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.